THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 59, No. 4, April 1980
Printed in U.S.A.

Recognition of Spoken Spelled Names for Directory Assistance Using Speaker-Independent Templates

By A. E. ROSENBERG, L. R. RABINER, and J. G. WILPON

(Manuscript received September 26, 1979)

In a recent paper, Rosenberg and Schmidt demonstrated the applicability of a speaker-trained, isolated word speech recognizer to the problem of automatic directory assistance. Input to the system was in the form of a string of letters which spelled the last name and initials of an individual for whom a directory listing was required. Rosenberg and Schmidt found that, even though the recognition rate for individual letters was rather low (approximately 80 percent), the rate at which the correct directory listing was found was higher (approximately 95 percent). In this paper, we extend these results to include the case of speaker-independent recognition of letters. We show that overall performance in the speaker-independent mode is comparable to performance in a speaker-dependent mode and examine various factors important for operation in a speaker-independent mode, such as characteristics of the reference templates, choice of decision rule, and threshold parameters. For the most part, the overall system is remarkably robust to the parameters of the recognizer. For the best choice of these parameters, a 95-percent correct string rate is obtained, comparable to the performance in a speaker-dependent mode.

I. INTRODUCTION

In the past few years, a great deal has been learned about the processes of automatic speech recognition. As a result, it is now practical to implement isolated word recognizers which achieve high accuracy for a variety of talkers for which the system is trained.[1] For the case of word recognizers that are speaker-independent, we are only now learning how to reliably implement such systems based on statistical characterizations of the variability in speaking the words of the

vocabulary.[2–5] Because of the success of these recognition efforts, work has been proceeding on task-oriented applications of isolated word recognition.

One such example of a task-oriented recognizer is the directory assistance system proposed by Rosenberg and Schmidt.[6] For this system, an isolated word recognizer is linked to a post-processing directory search algorithm to find a name in the directory that best matches the recognizer output (which is, of course, an estimate of the spoken name). A block diagram of this system is given in Fig. 1. The input to the system is a string of isolated words consisting of the letters of the last name (up to six letters), followed by the command word STOP, followed by one or more initials, and a final STOP (to delimit the end of the string). The isolated word recognizer provides a set of candidates for each spoken word (as shown in Fig. 1 for the name LEE K) ordered by recognition distance. The matrix of recognition candidates is used by a post-processor directory search procedure which queries a directory of names to find a best match to the candidate string.

Rosenberg and Schmidt evaluated the performance of the system of Fig. 1 using a speaker-trained recognizer trained to each of 10 talkers.[6] Although the acoustic recognition rate of the system on the letters was moderate (on the order of 70 to 80 percent), the recognition rate on the names was reasonably high (94 to 96 percent) due to the powerful contextual constraints imposed by the directory of names. Since the results of the earlier study were so encouraging, the system has been evaluated using a speaker-independent word recognizer in place of the speaker-trained recognizer. It is the purpose of this paper to discuss the results of this modified system and to compare its performance to the system where speaker-dependent templates are used. We will also discuss the effect on system performance of some experimental variables of the recognizer such as decision and threshold parameters and method of template construction.

Before presenting a discussion of the research which was carried

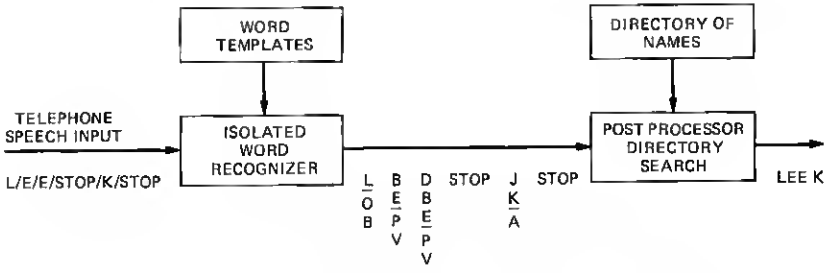

Fig. 1—Overall block diagram of the directory assistance system.

out, a few comments are worth making on the broad outlook of this type of recognition system. The first comment concerns the use of isolated words (letters) to spell a name rather than a connected string of letters. Although connected input is more natural than isolated input, it is not currently feasible to accurately recognize connected letters. As such, the isolated word format is the only current practical choice for the input to the system. The next comment concerns the issue of making the system speaker independent. In order for a system like the proposed one to be practical, it must operate in a speaker-independent manner, since it is not feasible to train, or store, templates for each possible user. Thus the work described here is not only a forward step, it is a necessary step for the system to be useful for the intended application. The final comment concerns the post-processing directory search algorithm. As mentioned earlier, the search procedure used here is a form of backtracking algorithm. Recently, Aldefeld et al have proposed a more efficient, more accurate minimum distance searching algorithm for use in the directory assistance system.[7] Although this new search method effectively replaces the one used here (and is currently being studied in this application), the results to be presented here reflect the broad interactions between the recognizer and the post-processor, and are essentially independent of the details of the search method.

The outline of this paper is as follows. In Section II we review the operation of the major components of the directory assistance system, i.e., the word recognizer and the post-processing search method. Section III describes the techniques used to evaluate system performance and to study the effects of various recognition parameters. Results of the investigations are given in Section IV, and in Sections V and VI these results are discussed.

II. OVERALL SYSTEM DESCRIPTION

We again refer to the block diagram of the overall spoken input, directory assistance system as shown in Fig. 1. Access to the system is via an ordinary telephone handset over a dialed-up connection through the local PBX. Upon receipt of an audible cue, the customer utters a string of letters which spell the last name and initials of an individual as listed in a telephone directory. The 18,000-entry Bell Laboratories directory is used in this implementation. The individual letters must be spoken distinctly and in an isolated manner (i.e., separated by intervals of at least 100 ms). The letters of the last name are spelled first, followed by a "stop" command followed by the initials, and a final "stop" command. Last-name lengths can be truncated to a specified maximum number of letters. Digits are included in the vocabulary to enable disambiguating information to be supplied when identical

names are encountered. Such information can be in the form of a four-digit organization number for Bell Laboratories directory listings.

The acoustic string of spelled letters is digitized and presented to a speech recognizer (an isolated word recognizer) which outputs candidate name strings to be searched in the telephone directory. For an actual implementation of the system, directory listing information corresponding to matching entries are read back to the customers via a conventional voice response unit. For the experimental evaluation described in this paper, the vocabulary consisted of only the letters of the alphabet, and the voice response system played no role.

2.1 Generating candidate strings

A detailed description of the process for generating candidate strings is given in Fig. 2, which shows the output and intermediate stages corresponding to a single uttered word at the input. The basic components of the system are as follows. First, there is an isolated word recognizer whose operation is based on reduction of the digitized speech signal of an input utterance to sets of eighth-order linear prediction coefficients (LPC) spaced at 15-ms intervals throughout the utterance and a dynamic programming time alignment and matching procedure for comparing the input utterance with a set of word reference templates. The details of the word recognizer have been adequately described in the references[8] and will be omitted here.

The speaker-independent reference template store consists of NT templates for each of the NW words in the vocabulary. NT is set to 12 and NW equals 39 for the vocabulary shown in Table I.

The speaker-independent reference templates were derived in the

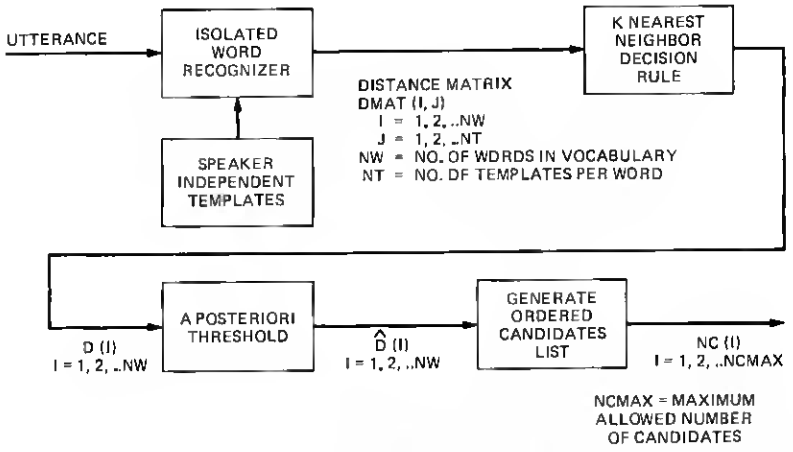

Fig. 2—Block diagram of the process for generating candidate words.

Table I—Vocabulary

1. A	14. N	27. Stop
2. B	15. O	28. Error
3. C	16. P	29. Repeat
4. D	17. Q	30. Zero
5. E	18. R	31. One
6. F	19. S	32. Two
7. G	20. T	33. Three
8. H	21. U	34. Four
9. I	22. V	35. Five
10. J	23. W	36. Six
11. K	24. X	37. Seven
12. L	25. Y	38. Eight
13. M	26. Z	39. Nine

following way. A population of 100 speakers, 50 male and 50 female, provided a single utterance for each word in the vocabulary shown in Table I. These utterances were carefully edited, digitized, and reduced to LPC templates. Using the dynamic programming time alignment and matching procedure, a measure of dissimilarity or distance was calculated for each pair of the 100 utterances for each word in the vocabulary. These pairwise distances were input to a statistical clustering process whose output was a classification of the 100 templates into a much smaller set of template groups or clusters. Each cluster can be represented either by the template at the cluster center or by a template which is the average of the templates included in the cluster. The basic principle underlying the use of this statistical clustering procedure is that if the 100-speaker population adequately represents a large fraction of the speaking population as a whole for this vocabulary, then the reduced set of clusters will also fairly represent this whole population. A detailed description of the statistical clustering procedures and an evaluation of recognizer performance using the resulting templates can be found in Refs. 3 to 5.

The output of the word recognizer is a matrix of distances or dissimilarity values for the comparison of the input word with each of the NW × NT reference templates.

The second step in the process is combining the distances by invoking some decision rule to provide a single distance figure for each word in the vocabulary. The rule that has been chosen is the K-nearest neighbor (KNN) rule,[3] in which the combined distance is the average of the K best (smallest) distances for each vocabulary word.

The following two steps are carried out to restrict the number of candidate words provided to the directory search. First, a distance rejection threshold is imposed to eliminate as candidate words all templates whose distances exceed the threshold. Second, the candidate words are ordered by their distances and no more than NCMAX of them are admitted.

The output of the recognizer is an ordered set of candidate words corresponding to the word uttered as input. An array of candidate letters corresponding to a string of uttered letters forming a name is provided to search the telephone directory in an attempt to find an entry which matches some combination of the candidate letters.

Based on the above discussion, we see that the major experimental variables in the recognizer that can affect the performance of the directory assistance system are:

(*i*) Speaker-dependent versus speaker-independent templates.

(*ii*) For speaker-independent templates, the method of template construction.

(*iii*) The KNN rule for recognition.

(*iv*) The distance rejection threshold level.

(*v*) The cutoff number of candidates for each letter (NCMAX).

Although each of the above variables has been studied in isolation (i.e., without context), it is important to understand its effects in a task-oriented application such as the directory assistance system. We investigate such effects in Section III.

2.2 Post-processor directory search method

The overall search and matching procedure is outlined in the simplified flowchart shown in Fig. 3. The directory search is a kind of iterative "backtracking" process. The directory is probed to find that entry which has the greatest consecutive number of letters starting with the first position which match the letters of the current candidate string. New candidate strings are formed by replacing the candidate letter in the left-most mismatch position of the last candidate string by the next best letter available in that position from the candidate letter array. If all the candidate letters in that position are exhausted, it is replaced by its best candidate letter and the preceding position is selected for candidate letter replacement. The search fails when the selected position is "backtracked" beyond the first position. The matching and search procedure allows for the possibility of no candidate letters at all in one or more positions by permitting a match to any letter in such positions. In addition, such "wild card" positions can be imposed one position at a time after all candidate strings provided by the recognizer have been exhausted. The details describing the directory search and match procedure are found in Ref. 6.

III. EVALUATION DESCRIPTION

Fourteen adult speakers (seven male and seven female) participated in an evaluation of the directory assistance system. The subjects were all native speakers of English and were unpaid volunteers. Seven of these speakers were included in the training set for the construction of

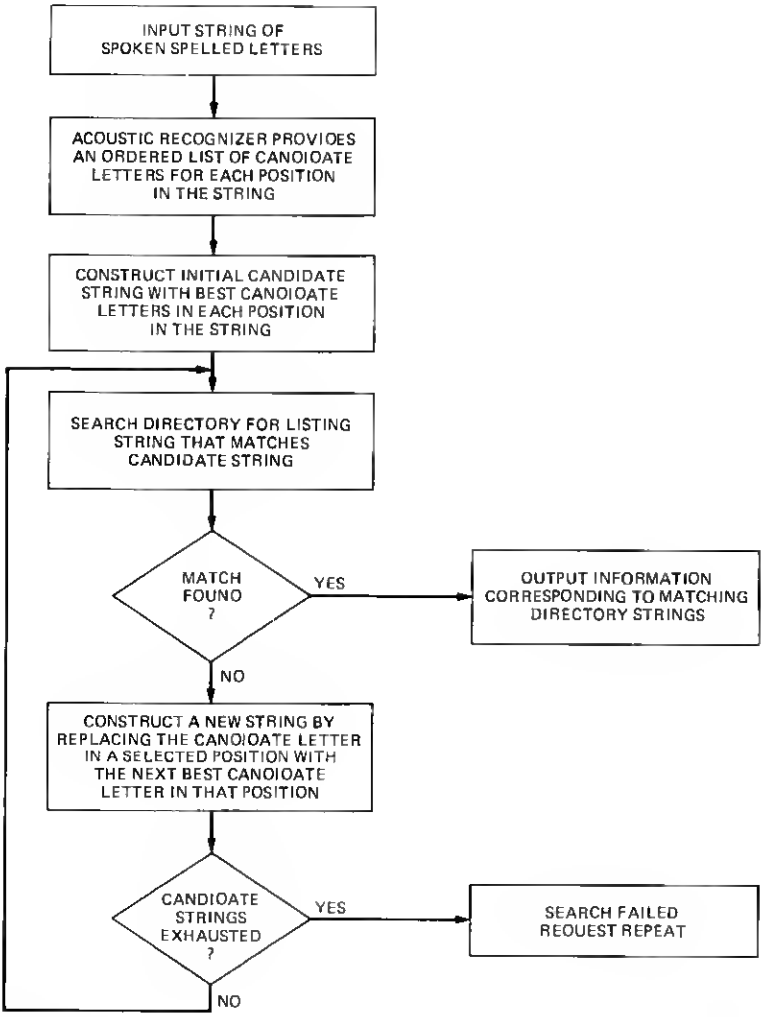

Fig. 3—Flow chart showing search and matching procedure for the directory searching post-processor.

the speaker-independent templates and ten speakers participated in the earlier experiment in which only speaker-dependent templates were used. For comparison, the speaker-dependent template condition was repeated in this evaluation. Those speakers who did not participate in the earlier experiment were required to establish two reference templates for each item in the vocabulary for the speaker-dependent condition. These were obtained in two separate sessions for each speaker lasting approximately 5 minutes each.

In the test sessions, each speaker spelled each of 50 names from a

list randomly selected from the Bell Laboratories directory. The list is shown in Table II and is the same one used in the earlier evaluation. In this experiment, last-name lengths are truncated to six letters. Also, the test utterances consist only of letters of the alphabet. Digits are not included, nor is "STOP," which is assumed to be recognized without error. Thus the experimental value of NW is 26. Most speakers required two sessions to complete the 50-name list. The talkers provided their utterances in a quiet computer laboratory, using an ordinary telephone handset over dialed-up telephone lines. The utterances were digitized and preprocessed on-line using a Data General Eclipse S-230 laboratory computer. Recognition and directory searching and matching were carried out off-line using the same facility. Typically, a set of 50 names took 2 hours for recognition and about 1 hour for directory search.

3.1 Experimental parameters

3.1.1 Reference templates

In addition to speaker-dependent templates, three types of speaker-independent templates were investigated. The first type of templates was produced by an interactive clustering technique in which an operator guided the clustering analysis. Such techniques are called supervised procedures. In this method, each cluster was represented by its minimax center.[3] The second and third types of templates

Table II—List of test names used in the evaluation. Letters in parentheses were omitted.

1. ZBOYAN A M	26. TINLEY M A
2. KRIEGER(R) G E	27. SHAEFF(ER) P A
3. ROHM B J	28. LIND G R
4. EPWORT(H) R	29. SHIPLE(Y) J W
5. LINDHA(RD) E A	30. CUCCO J A
6. BURNS J F	31. HOFER F R
7. RUDDOC(K)B	32. DUNBAR J J
8. SCHILL(O) R F	33. DUKE S D
9. GOOZH J L	34. WASSON R D
10. VIROST(EK) A M	35. HOOO A A
11. LENNON F W	36. MENGEL M R
12. VASHIS(HTA) P	37. RAVER D F
13. DUFFY G L	38. FULTZ K E
14. YAEGER J C	39. CAOWEL(L) K
15. CRAWFO(RD) C D	40. YOUHAS J M
16. WEEKS C G	41. VANBEN(THEM) J
17. AVEYAR(D) R L	42. OUNCAN(SON) J P
18. GRECO T J	43. SUMNER E E
19. MODARR(ESSI) A R	44. LAWREN(Z) D A
20. ERWIN W J	45. BLY J
21. LUM P S	46. NEWELL J A
22. SOOS N A	47. STAUBA(CH) W E
23. SKARIN R H	48. TATE B A
24. TENEYC(K) J H	49. ONDER J J
25. SACCO G A	50. SOLOMI(TA) K S

represented clusters by averaging the tokens in each cluster. The averaging was carried out using the autocorrelation coefficient representation of the templates.[5] The second type of templates was obtained from the supervised method, and the third type was unsupervised (i.e., obtained from a fully automatic clustering procedure). It is important that the performance of the system with the unsupervised templates be essentially comparable to that with supervised templates, since this case represents the most practical one for most applications.

As noted in the previous section, there were two templates per word for speaker-dependent templates and 12 templates per word for speaker-independent templates.

3.1.2 K-nearest neighbor decision rule for speaker-independent templates

It has already been noted, in the speaker-independent case, that the distances output by the recognizer for each template were combined into a single distance, for each word in the vocabulary, using the K-nearest neighbor rule[3] (KNN). Three values of K were used, namely, $K = 1$, 2, and 3.

The KNN rule is not an experimental variable for the speaker-dependent condition where there are just two templates per vocabulary item. For this condition, the minimum distance, nearest neighbor rule ($K = 1$) was applied.

3.1.3 Rejection threshold

The results will demonstrate that the performance of the "backtrack" procedure for searching and matching directory strings is sensitive to the number of candidate letters allowed per letter in that the search time increases rapidly as the number of candidate letters increases. One way to control the number of candidate letters is to impose a rejection threshold which admits only those candidate letters whose templates have distances less than the threshold. Nine *a posteriori* threshold values were examined in this experiment. In addition, the value of NCMAX was set to 10.

IV. RESULTS

4.1 Speaker-independent versus speaker-dependent performance

The performance of the system is best specified in terms of the number of spelled names matched correctly to names in the directory. The rate at which errors are made for this performance criterion is termed *string error rate*. Results for the speaker-independent condition are represented by Type 2 templates, obtained by averaging in a supervised procedure, using the KNN rule with $K = 2$. The error rates

shown are obtained by averaging the error rates for the three best rejection threshold conditions for each individual. This averaging procedure is invoked to provide more representative results, since there is a fair amount of fluctuation of string error rate as a function of threshold for some individuals in the region of optimum threshold. Individual string error rates for speaker-independent and speaker-dependent templates are shown connected in Fig. 4. The mean error rate for all talkers, shown by solid arrows, is 9.9 percent for the speaker-independent condition and 4.9 percent for the speaker-dependent condition. Individuals included in the training set for constructing

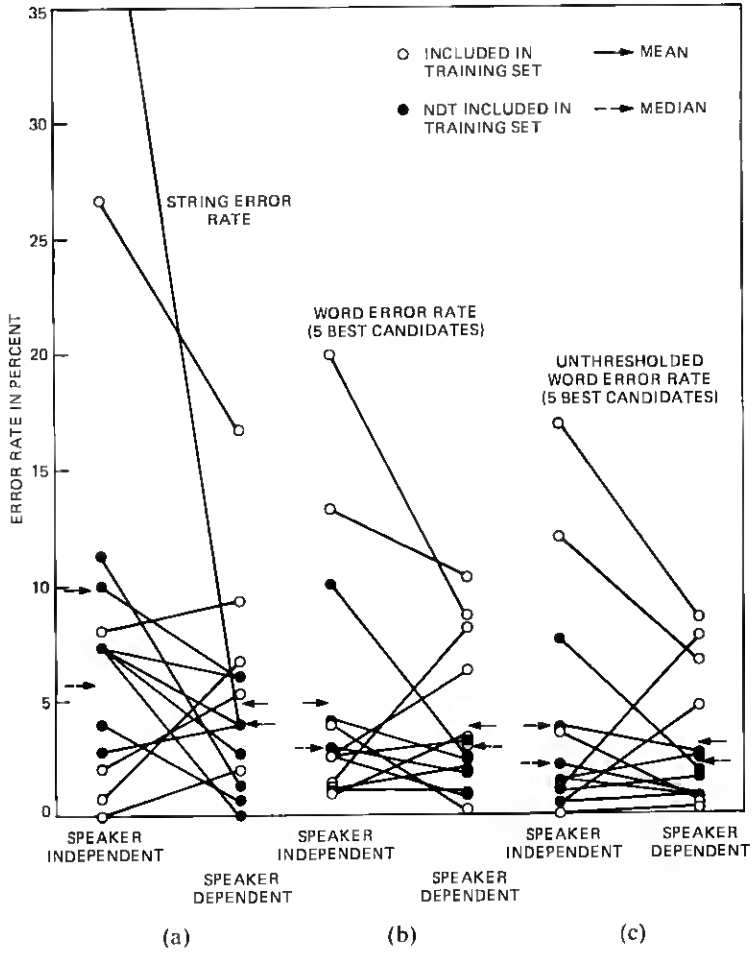

Fig. 4—Overall error rates for individual talkers shown for Type 2 speaker-independent templates and speaker-dependent templates. (a) String error rate. (b) Word error rate (five best candidates). (c) Word error rate (five best candidates) unthresholded.

the speaker-independent templates are distinguished by different symbols from those who are not. A statistical significance test, the Mann-Whitney U-Test,[9] failed to detect a significant difference (i.e., $U(7) = 22.5$, $p < 0.42$) between the performances of these two populations for either the speaker-independent or speaker-dependent condition. Of course, a difference in performance could only be anticipated for the speaker-independent population.

The Mann-Whitney U-Test was also invoked to detect significant differences between the speaker-independent and speaker-dependent performances. In this case, a marginally statistically significant difference at the 10-percent level ($U(14) = 62.5$) was found. The difference is apparently accounted for by the two worst individual rates for the speaker-independent condition. Curiously, these two individuals, with 26.7- and 44.0-percent error rates, were both included in the training set for template construction. Although their performances were also generally inferior for the speaker-dependent condition, only for the speaker-independent condition were their performances so radically separated from the rest of the population. This effect is discussed later in this paper. The effect of removing the extremes of the distribution (including these two individuals) from the population by taking the median of the data (rather than the mean) is indicated by the arrows in Fig. 4. Difference in performance between speaker-independent and speaker-dependent populations is considerably narrowed with error rates of 5.7 and 4.0 percent, respectively. There is no detectable statistically significant difference ($U(12) = 49.0$, $p > 0.1$).

Error rates for individual letters in the name strings are denoted word error rates. These error rates characterize the performance of the acoustical recognition components of the system in contrast to string error rates which characterize whole system performance. Word error rates (i.e., the rate at which the correct word is not among the N best candidates for recognition), in the absence of a rejection threshold, are shown plotted in Fig. 5a for the speaker-independent condition and 5b for the speaker-dependent condition. (The speaker-independent condition uses the same template type and KNN rule parameter as observed in Fig. 4a.) Word error rate is also shown plotted as a function of N best candidates, or candidates with the N smallest distances, for the mean and median across the 14 talkers as well as the best and worst individual talkers. The conventional performance measure for recognition accuracy is the $N = 1$ rate, the rate for which the best candidate is incorrect. By this criterion, the recognizer performs at a mean error rate of 23.5 percent for speaker-independent templates and 19.0 percent for speaker-dependent templates. (In the earlier experiment using speaker-dependent templates, the median word error rate over 10 talkers was 20.5 percent.) Performance improves with increas-

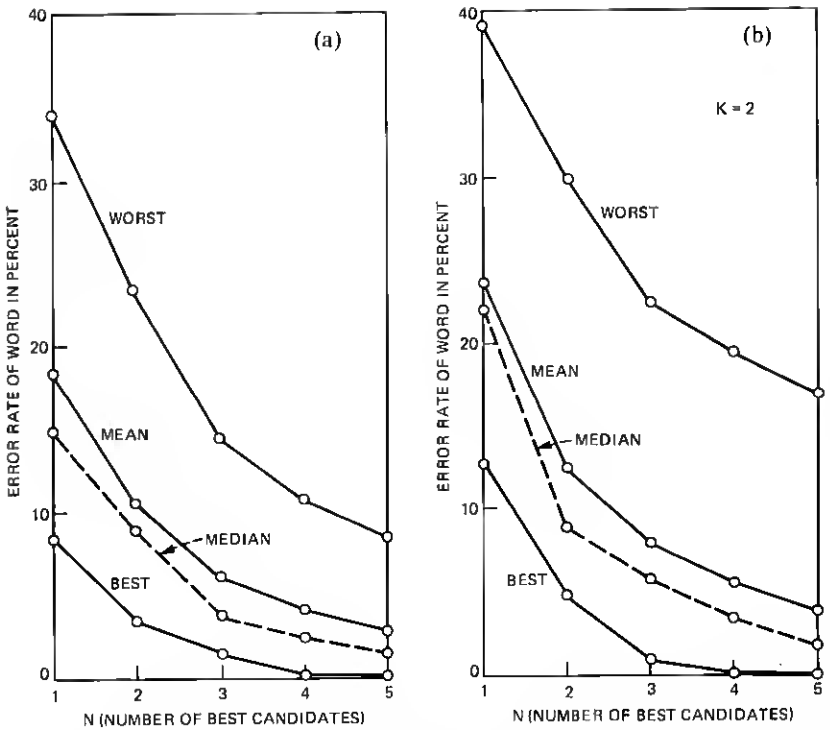

Fig. 5—Word error rates as a function of number of best candidates showing the mean and median over all the talkers as well as the best and worst individual talker. (a) Speaker dependent. (b) Type 2 speaker independent.

ing number of best candidates, although differential improvement in performance decreases.

For five best candidates, the mean error rate is 3.9 percent for speaker-independent templates and 2.9 percent for speaker-dependent templates. It was pointed out in the earlier experiment[6] that the five-best-candidate performance figure is more meaningful for this system than the conventional, best-candidate recognizer performance figure. This is because candidate name strings are constructed from a list of best candidate letters (as many as 10 in this experiment) for each position in the string. In fact, in the earlier experiment it was shown that word error rate for five best candidates is a reasonably good predictor for string error rate.

In Fig. 4, along with individual string error rates shown in 4a, individual word error rates (five best candidates) are plotted in 4b and 4c. The rates shown in Fig. 4b are associated with the same threshold conditions for the individual string error rates shown in 4a. Individual error rates with no threshold imposed are shown in 4c. Except for the

two worst individual performances, word error rates for five best candidates are seen to be comparable to the string error rates. They do, however, exhibit less variability and occupy a smaller range than the string error rate. Such relative statistical stability is to be expected, since just a single word error out of some six or eight words in a string will result in a string error.

4.2 Effect of template type

To characterize the system performance for speaker-independent templates in Fig. 4, we chose templates that are constructed by averaging autocorrelation coefficients in a supervised procedure. In fact, there are no significant differences in string error performance among the three speaker-independent template types investigated in this experiment, as shown in Fig. 6 (values of U(14) range from 77 to 91.5). Shown plotted are overall string error rates, means, and medians, for each of the three speaker-independent template types, and for speaker-dependent templates. These were obtained in the same way as the data displayed in Fig. 4 (i.e., by averaging the three best threshold conditions for each individual talker). Also plotted in this figure are mean word error rates (five best candidates). For the word error rate criterion, it is seen that Type 2 templates have a slight

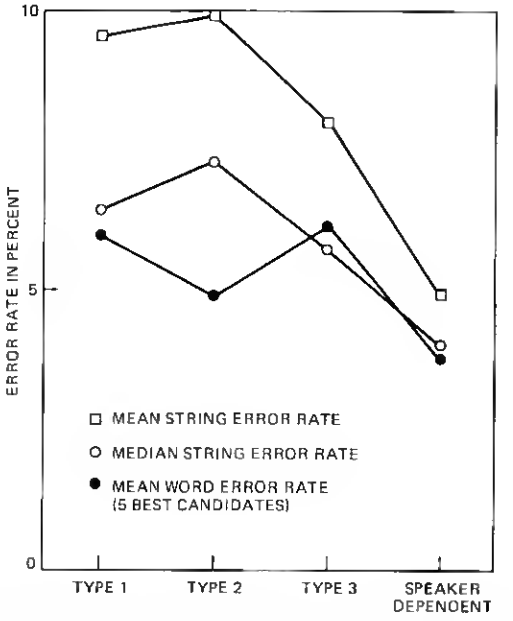

Fig. 6—Overall string error rate and word error rate (five best candidates) for three speaker-independent template types and speaker-dependent templates.

advantage. This is why they were chosen to represent speaker-independent templates in the overall results even though the resulting string error rate is slightly worse than for the other types as seen in the figure. However, the Mann-Whitney Test failed to detect any significant differences (significance levels all greater than 10 percent) among the speaker-independent template types for either string error rate or word error rate (five best candidates). Speaker-dependent templates, though, do have a statistically small but significant advantage over any of the speaker-independent types ($U(14) \approx 65$, corresponding to $p < 0.1$).

Significant differences among speaker-independent template types do show up in one measurement. Word error rates for each template type are shown plotted as a function of number of best candidates in Fig. 7. Statistically significant differences ($U(14) < 52.5$, $p < 0.05$) among the template types are found for $N = 1$, word error rate (best candidate) with the best performance for Type 2 templates. For N greater than 1, there are no statistically significant differences ($U(14) > 67.5$). Since the performance of the search algorithm is relatively insensitive to the $N = 1$ word error rate, this result has little impact.

We conclude that, for the purposes of this system, there is little choice among the three speaker-independent template types and only a marginal advantage for speaker-dependent templates.

4.3 Effect of KNN rule parameter

The effect of varying the K parameter for the KNN decision rule is shown in Fig. 8. All three experimental values of K were observed only for Type 1 templates. Mean string error rate, word error rate (best candidate), and word error rate (five best candidates) over all individuals are shown plotted as a function of K in Fig. 8a. In an earlier experiment [3] it was shown that setting the K parameter to 2 provided best performance for reference data containing 12 templates per word. Although examination of Fig. 8a shows no clear trend for that value, if the two worst individual performances are removed, trends become apparent. Figure 8b plots the same variables as a function of K as shown in 8a for the 12 best individual performances, while the two worst performances are plotted in Fig. 8c. It seems clear that $K = 2$ or 3 is preferred for the best performances while $K = 1$ has a strong advantage for the worst performances.

A possible explanation is the following. The worst talkers may have few reliable reference templates, perhaps only one per word. In such a situation, it can only be harmful to average the distance of a reliable template with others that are not. Thus, the predicted stabilizing effect of using the KNN rule with $K = 2$ or 3 for the 12-template-per-word vocabulary can only be attained when there are a sufficient number of

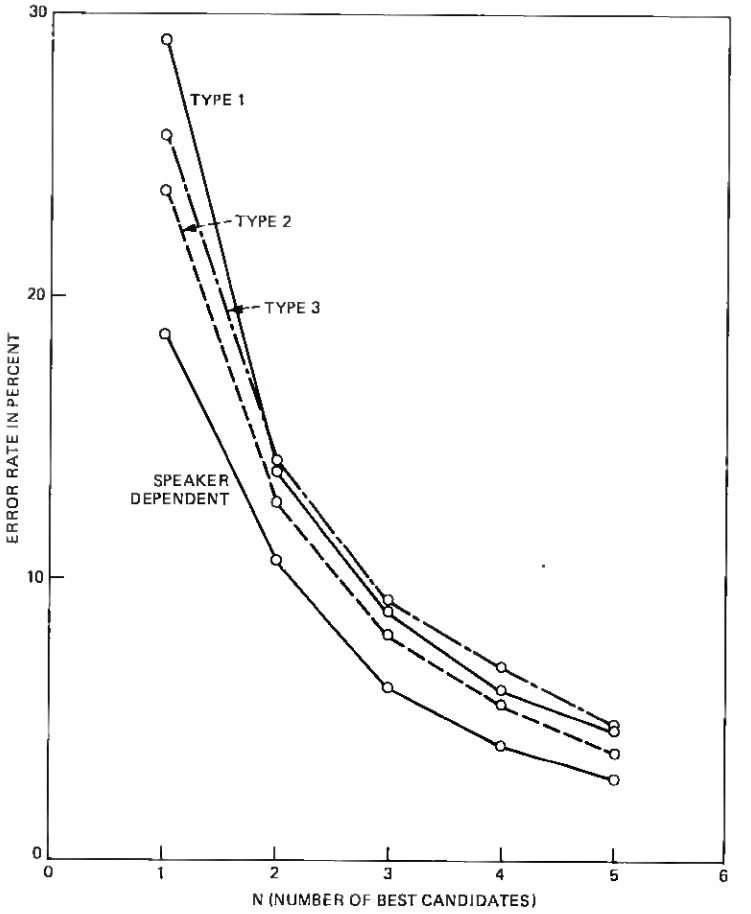

Fig. 7—Mean word error rate (unthresholded) as a function of number of best candidates for three speaker-independent template types and speaker-dependent templates.

reliable templates per word. For such talkers, it would be advantageous to try to estimate the value of K that gives the best results on a calibration name, and then use this value of K for the system. This procedure is cumbersome and hampers the use of the system.

The effect of selecting the best K value for each individual talker is summarized in Table III, where means and medians of string error rates and word error rates are shown. Speaker-dependent rates are shown for comparison. It can be seen that selecting the K value which provides the best string error rate for each talker yields mean and median string error rate of 6.6 and 4.7 percent, respectively. This represents a 25- or 30-percent improvement in string error rate over the condition when the compromise K value of 2 is used. The corre-

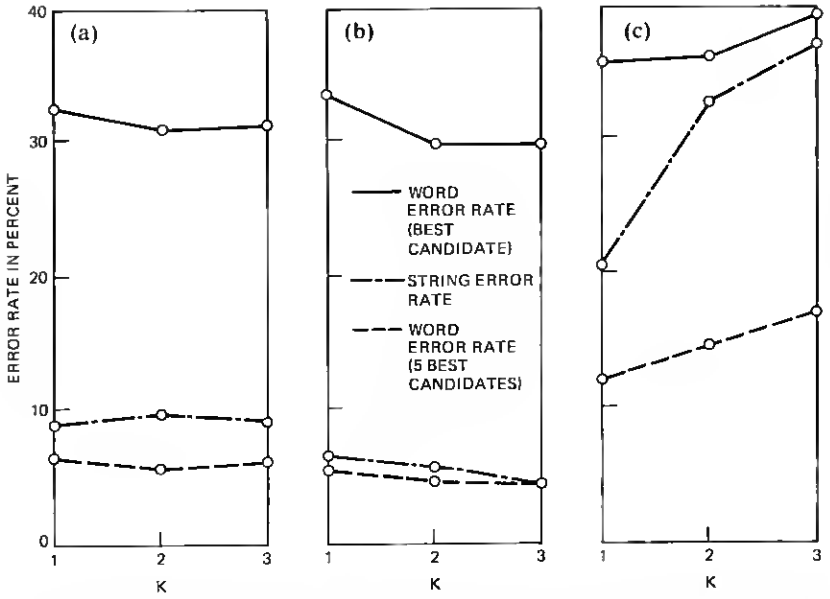

Fig. 8—Mean word error rate (best candidate), word error rate (five best candidates), and string error rate as a function of KNN decision rule parameter. (a) All talkers. (b) Twelve best talkers. (c) Two worst talkers.

sponding improvement in word error rate (five best candidates) is about 10 percent, while the improvement in word error rate (best candidate) is negligible. It seems that string error rate is fairly sensitive to choice of K value. As implied in Fig. 8, the best choice of K is three for the 12 best talkers and one for the two worst talkers.

4.4 Effect of threshold variation

As mentioned earlier, the rejection threshold, which is applied following the KNN decision rule, has the effect of limiting the number

Table III—Overall error rates obtained in the evaluation

	$K = 2$	Best K	Speaker Dependent
String Error Rate			
Mean	9.5	6.6	4.9
Median	6.4	4.7	4.0
Word Error Rate			
Best candidate			
Mean	30.8	30.6	19.3
Median	30.2	29.7	16.2
Best 5 candidates			
Mean	5.7	5.3	3.8
Median	4.7	4.2	2.8

of candidate letters available for constructing candidate strings. The threshold is a sensitive variable affecting system performance, since if too few candidate letters are admitted there is a risk of excluding the correct string as a match, while if too many candidate letters are admitted there is a risk of obtaining an incorrect match before the correct string is generated. This effect is shown in Fig. 9, where mean and median string error rates are plotted as a function of threshold. Also shown plotted as a function of threshold are mean word error rate (five best candidates) and mean number of candidates admitted per word. Type 2 templates are used with KNN rule value (K) set to 2. Both the mean and median string error rates have clear minima while word error rate approaches an asymptotic minimum at 4.0 percent, which is very nearly the unthresholded value. The mean number of candidates admitted per word monotonically increases as a function of threshold. It appears that an optimum uniform threshold setting is

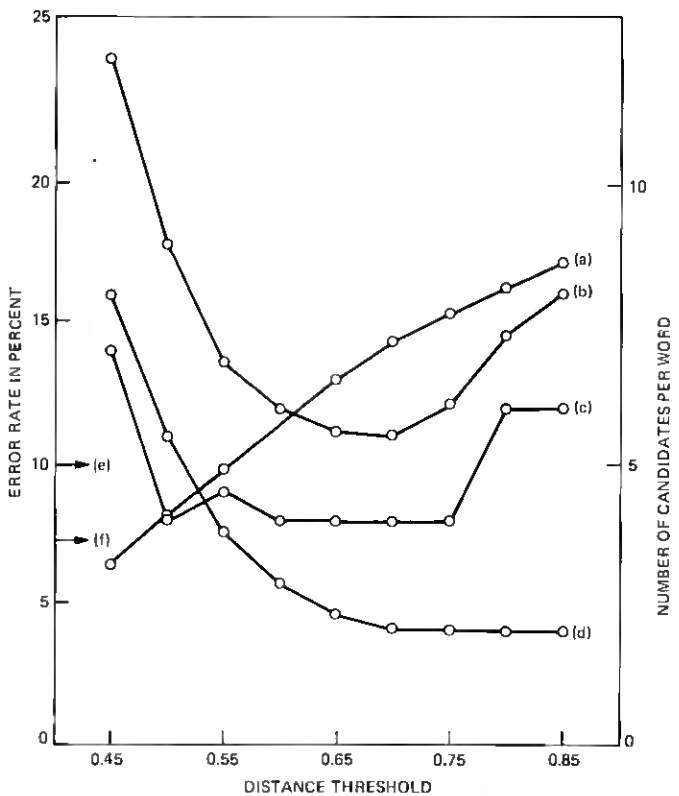

Fig. 9–Effect on performance of distance rejection threshold. (a) Mean number of candidates per word input. (b) Mean string error rate. (c) Median string error rate. (d) Mean word error rate (five best candidates). (e) Mean and (f) median string error rate using best individual talker threshold settings.

found just at the point where word error rate (five best candidates) is approaching a minimum and the mean number of candidates admitted is 6 or 7. Any further threshold relaxation decreases string error performance. For comparison, the mean and median string error rates for best individual threshold conditions are shown by the arrows along the ordinate.

4.5 Recognizer confusions

Additional insight into the recognition results can be gained by examining the confusions made by the recognizer among the vocabulary words. A confusion matrix is shown in Table IV, which provides the mean frequency of recognition for each specified letter over all the trials of all the talkers. A letter is defined as "recognized" if it is ranked better than the specified letter in the candidate list. (The specified letter is not recognized only when it is rejected from the candidate list.) Recognition frequencies less than 5 percent are omitted to simplify the matrix. This matrix is tabulated from Type 2 templates with KNN rule value set to 2. A similar matrix was presented in the original experiment for speaker-dependent template data.[5] The present results are not materially different from the earlier results. As expected, there are a large number of confusions among the "BDE ... " family of letters, as well as "AJK," "FS," "MN," and "IY." Only rarely are the confusions symmetric. For example, "A" is recognized as "K" with a frequency of 21 percent, but there is no significant confusion of "K"

Table IV—Letter recognition confusion matrix, The numbers in parentheses are percent distribution frequencies for individual letters.

	Frequency of Recognition as a Function of Specified Letter Recognized As																		
Specified As	B	D	E	G	P	T	V	Z	C	A	J	K	I	Y	F	S	X	M	N
B (2.5)	30	14	7	5	14	6	12	5											
D (4.7)	7	34	5	11	14	11	7	7											
E (9.6)	5	13	61	6	7														
G (2.8)		7	6	62	8	6													
P (1.6)		11		12	36	14	12	7											
T (2.5)		8		11	13	32	6	9	8										
V (1.4)	5	7		6	6	5	38	21	7										
Z (1.0)				7			9	59	12										
C (4.1)								13	79										
A (9.9)		6		5		6	5	5	6	34	8	15							
J (4.9)				5				7			74	6							
K (1.9)				6							12	78							
I (3.0)													85	13					
Y (2.2)														100					
F (3.3)															85	14			
S (5.8)															10	85	5		
M (3.0)																		63	25
N (6.3)																		5	75

for "A." The confusions tabulated in this matrix well illustrate the difficulties associated with this vocabulary for recognition.

V. SUMMARY AND DISCUSSION

The principal results of this experiment are twofold. First, context constraint in the form of spelled names as listed in a telephone directory has a powerful correcting influence on the performance of an automatic directory assistance system with a vocabulary composed of letters of the alphabet. This result is a confirmation of the results of an earlier experiment and is shown by a comparison of error rates for individual words of the order of 30 percent, with error rates for name strings of the order of 5 percent. Second, only a slightly significant degradation in performance is associated with using speaker-independent templates instead of speaker-dependent templates. In addition, we have seen that, for the speaker-independent case, the performance for individuals included in the training set is essentially the same as for those who were not in the training set.

The results suggest in a qualitative way that there is more variability in performance using speaker-independent templates than with speaker-dependent templates. A particularly conspicuous example is the performance of one of the two worst talkers, whose error rates are markedly worse than the mean over all talkers for speaker-independent templates but close to the mean for speaker-dependent templates (see Fig. 4).

A possible explanation for the apparent greater variability and general degradation in performance for speaker-independent templates compared with speaker-dependent templates lies in the consideration of what we might call catastrophic errors. A catastrophic word error is defined as an error in which the correct word is ranked worse than tenth in the list of word candidates. Such a ranking removes the correct letter completely from the candidates available for searching the directory for a matching string. As a result, the chances of obtaining a string error are greatly increased, since the "wild card" procedure mentioned at the end of Section II can handle at most one such error.

The number of catastrophic word errors has been tabulated for the best individual threshold conditions for speaker-dependent templates and for template Type 3 and $K = 3$ speaker-independent templates. Percentages of 1.6 of the total number of matches for speaker-dependent templates and 2.8 for speaker-independent templates were catastrophic errors. Moreover, in rare but striking instances catastrophic errors are distributed quite nonuniformly across the vocabulary. For example, for the worst speaker using speaker-independent templates, 75 percent of the "A's," 37 percent of the "R's," 34 percent of the "N's," and 26 percent of the "O's" resulted in catastrophic errors with

just a few additional errors scattered over the rest of the vocabulary. Such repeatedly poor performance on these letters, not surprisingly, resulted in poor string error performance.

There are two possible sources of catastrophic word errors. First, the test utterance may be botched or severely degraded by some disturbance to the extent that it does not resemble any of the correct reference templates. The second possibility is that the correct reference templates may not all be characteristic of the speaker's normal utterances. It seems likely that occasional catastrophic errors should result from the first condition, but that repeated catastrophic errors probably arise from the second condition.

Although the second condition might occur for either speaker-dependent or speaker-independent templates, there seems to be a distinct, although perhaps rare, possibility that none of the 12 speaker-independent templates for a particular word in the vocabulary are characteristic of a given speaker's utterances. The result would be the repeated catastrophic errors that have been observed for the worst performing speaker.

Although Rabiner and Wilpon[5] have demonstrated distinct differences in performance among the three speaker-independent template types used in this experiment, with the most favorable performance associated with Type 2 templates, no significant differences were apparent in the results of this experiment. The differences observed by Rabiner and Wilpon are associated with word error rate (best candidate). However, in this experiment, string error performance is relatively insensitive to the best candidate word for recognition, depending only on the presence of the correct candidate in a list of up to 10 candidates (as seen in Fig. 6). As shown earlier, the performance of this system is more closely related to the word error rate for five best candidates, and this score is relatively insensitive to template type.

It has been shown that use of the K-nearest neighbor decision rule with K set to 2 or 3 produces a clear advantage in performance over the conventional nearest-neighbor decision rule (K=1) with one important caveat. This is that there must be a sufficient number of reliable templates among the 12 for each word. We have shown that this condition did not hold for the two worst talkers, in which case K should be set to 1.

The necessity for imposing a threshold distance to limit the number of candidate words supplied for the search for a matching directory entry has been demonstrated in Section 4.4. When the threshold is relaxed beyond a certain optimum setting, allowing more candidate words, string error rate actually increases. In addition, increasing the number of candidate words lengthens the overall search time. Although it is possible to set a uniform threshold for all talkers, there is a

sensitive dependence on threshold for individual talkers, and performance can be improved by setting the threshold individually.

We have also discussed a possible adverse effect of imposing a maximum limit on the number of candidate words. This is the occurrence of repeated "catastrophic" word errors in which the correct candidate is excluded from the search. The result is a poor string error rate. Since this experiment was completed, Aldefeld et al[7] have demonstrated an improved search procedure which does not require that the number of candidate words be limited and results in greatly improved string accuracy of the order of 98 percent. It makes full use of the distances associated with each candidate word and demonstrates that the best match is that string in the directory whose total distance is a minimum. The improved search procedure is also considerably more efficient in operation. This efficiency is accomplished by partitioning the directory according to equivalence classes based on acoustical similarity using the information provided by the confusion matrix shown in Table IV.

VI. CONCLUSION

We have confirmed the results of an earlier experiment demonstrating the powerful effects of spelling context on the automatic recognition of spoken letters of the alphabet with specific application to the retrieval of telephone directory entries, showing that the system can operate in a speaker-independent mode nearly as well as the earlier investigated speaker-dependent mode. Several experimental conditions affecting performance associated with using speaker-independent templates have been examined, including template construction technique, decision rule, and threshold. Under optimum conditions, mean string error rate is approximately 5 percent for speaker-independent templates.

VII. ACKNOWLEDGMENT

The authors acknowledge the meticulous and thoughtful comments of an anonymous B.S.T.J. reviewer, which hopefully improved the clarity and usefulness of this paper.

REFERENCES

1. T. B. Martin, "Practical Applications of Voice Input to Machines," Proc. IEEE, *64* (April 1976), pp. 487–501.
2. L. R. Rabiner, "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words," IEEE Trans. on Acoustics, Speech, and Signal Proc., *ASSP-26* (1978) pp. 34–42.
3. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," IEEE Trans. on Acoustics, Speech, and Signal Proc., *ASSP-27* (1979), pp. 336–349.

4. S. E. Levinson, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "Interactive Clustering Techniques for Selecting Speaker Independent Reference Templates for Isolated Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Proc., *ASSP-27* (1979), pp. 134–141.
5. L. R. Rabiner and J. G. Wilpon, "Considerations in Applying Clustering Techniques to Speaker Independent Word Recognition," J. Acoust. Soc. Am., *66*, No. 3 (Sept. 1979), pp. 663–673.
6. A. E. Rosenberg and C. E. Schmidt, "Automatic Recognition of Spoken Spelled Names for Obtaining Directory Listings," B.S.T.J., *58*, No. 8 (October 1979), pp. 1797–1823.
7. B. Aldefeld, S. E. Levinson, and T. G. Szymanski, "A Minimum-Distance Search Technique and its Application to Automatic Directory Assistance," unpublished work.
8. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. on Acoustics, Speech, and Signal Proc., *ASSP-23* (1975), pp. 67–72.
9. S. Siegel, *Nonparametric Statistics for the Behavioral Sciences*, New York: McGraw-Hill, 1956, pp. 116–127.